The NCBI Genomic Discovery System

Using Integrated Data from Multiple Genomes at NCBI to Accelerate Biological Discovery

Eric W. Sayers, Ph.D.
National Center for Biotechnology Information
Bethesda, MD USA

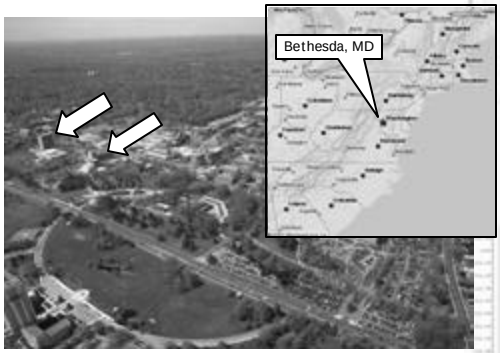
HGM2004
5 April 2004
17.00 – 19.00

This Evening's Menu...

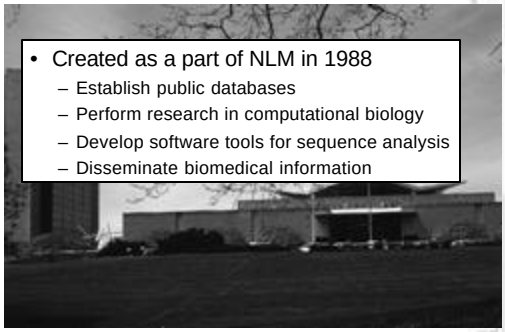
Gene of the Day: thyroid peroxidase

- Entrez Gene and RefSeq
  - Exploring evidence for gene annotations
- Beyond RefSeq
  - What if my species isn't in RefSeq?
- Sequence Polymorphisms
  - Finding SNPs and phenotypes
- NCBI Map Viewer
  - View and retrieve annotations and associated data
- Finding Gene Homologs
  - Find homologs using Entrez, Map Viewer, and BLAST
- Example Problems
  - Searching for markers and exploring disease alleles

The National Institutes of Health



The National Center for Biotechnology Information

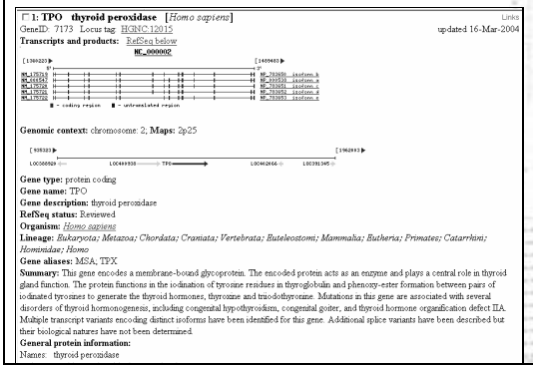


- Created as a part of NLM in 1988
  - Establish public databases
  - Perform research in computational biology
  - Develop software tools for sequence analysis
  - Disseminate biomedical information

Entrez Gene and RefSeq

- Entrez Gene is a new database of gene loci
- Entrez Gene includes all organisms in LocusLink plus all other organisms in RefSeq
- Entrez Gene is an Entrez database with full searching and linking capabilities
- Entrez Gene will ultimately replace LocusLink

Gene Record for Human TPO



1: **TPO thyroid peroxidase** [*Homo sapiens*] Links

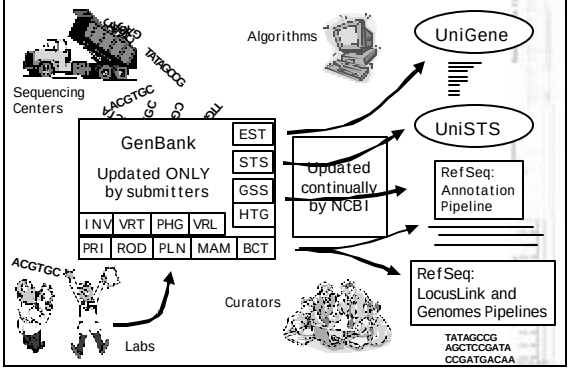
GeneID: 7173 Locus tag: HGNC:12015 updated 16-Mar-2004

Transcripts and products: RefSeq below

Genomic context: chromosome: 2; **Maps:** 2p25

Gene type: protein coding
Gene name: TPO
Gene description: thyroid peroxidase
RefSeq status: Reviewed
Organism: *Homo sapiens*
Lineage: *Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo*
Gene aliases: MSA, TPX
Summary: This gene encodes a membrane-bound glycoprotein. The encoded protein acts as an enzyme and plays a central role in thyroid gland function. The protein functions in the iodination of tyrosine residues in thyroglobulin and phenoxy-ester formation between pairs of iodinated tyrosines to generate the thyroid hormones, thyroxine and triiodothyronine. Mutations in this gene are associated with several disorders of thyroid hormonogenesis, including congenital hypothyroidism, congenital goiter, and thyroid hormone organification defect IIA. Multiple transcript variants encoding distinct isoforms have been identified for this gene. Additional splice variants have been described but their biological natures have not been determined.
General protein information:
Names: thyroid peroxidase

Primary vs. Derivative Databases



RefSeq Benefits

- Non-redundant
- Explicitly linked nucleotide and protein sequences
- Updates to reflect current sequence data and biology
- Data validation
- Format consistency
- Distinct accession series
- Stewardship by NCBI staff and collaborators

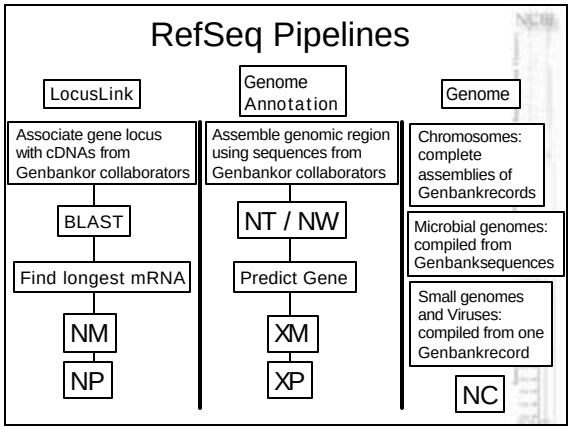
RefSeq Accession Numbers

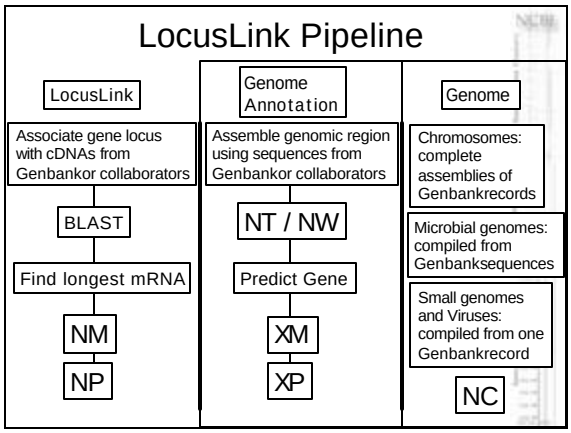
Curated Records

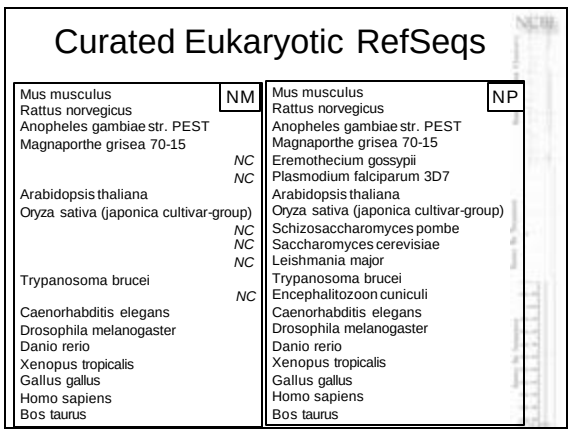
NC_123455	**Complete genomic sequence (chromosome)**
NG_123456	**Incomplete genomic sequence**
NM_123456	**mRNA**
NP_123456	**Protein derived from NM**
NR_123456	**Non-coding RNA**

Model Records

NT_123456	**Assembly of BAC data**
NW_123456	**Assembly of WGS data**
NZ_ABCD12345678	**Collection of WGS data**
XM_123456	**mRNA**
XP_123456	**Protein derived from XM**
ZP_123456	**Protein derived from NZ**
XR_123456	**Non-coding RNA**


RefSeq Pipelines
LocusLink
Associate gene locus with cDNAs from Genbankor collaborators
BLAST
Find longest mRNA
NM
NP
Genome Annotation
Assemble genomic region using sequences from Genbankor collaborators
NT / NW
Predict Gene
XM
XP
Genome
Chromosomes: complete assemblies of Genbankrecords
Microbial genomes: compiled from Genbanksequences
Small genomes and Viruses: compiled from one Genbankrecord
NC

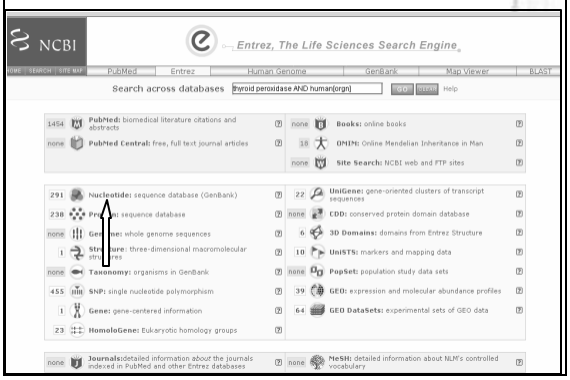

LocusLink Pipeline
LocusLink
Associate gene locus with cDNAs from Genbankor collaborators
BLAST
Find longest mRNA
NM
NP
Genome Annotation
Assemble genomic region using sequences from Genbankor collaborators
NT / NW
Predict Gene
XM
XP
Genome
Chromosomes: complete assemblies of Genbankrecords
Microbial genomes: compiled from Genbanksequences
Small genomes and Viruses: compiled from one Genbankrecord
NC


Curated Eukaryotic RefSeqs
NM
Mus musculus
Rattus norvegicus
Anopheles gambiae str. PEST
Magnaporthe grisea 70-15
NC
NC
Arabidopsis thaliana
Oryza sativa (japonica cultivar-group)
NC
NC
NC
Trypanosoma brucei
NC
Caenorhabditis elegans
Drosophila melanogaster
Danio rerio
Xenopus tropicalis
Gallus gallus
Homo sapiens
Bos taurus
NP
Mus musculus
Rattus norvegicus
Anopheles gambiae str. PEST
Magnaporthe grisea 70-15
Eremothecium gossypii
Plasmodium falciparum 3D7
Arabidopsis thaliana
Oryza sativa (japonica cultivar-group)
Schizosaccharomyces pombe
Saccharomyces cerevisiae
Leishmania major
Trypanosoma brucei
Encephalitozoon cuniculi
Caenorhabditis elegans
Drosophila melanogaster
Danio rerio
Xenopus tropicalis
Gallus gallus
Homo sapiens
Bos taurus

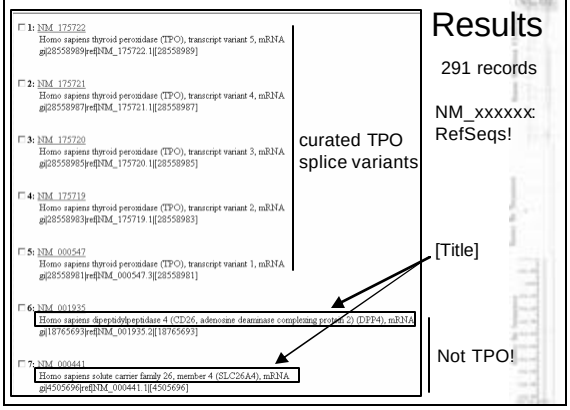
Finding Primary Sequences

- Search Entrez Nucleotide
  - 99.3% GenBank (primary data)
  - 0.7% RefSeq (curated data)
- Typical starting query:
  - `thyroid peroxidase AND human[orgn]`
- Useful search terms in [Properties]:
  - srcdb : Source database (srcdb genbank)
  - gbdiv : GenBank division (gbdiv est)
  - biomol : Biomolecule type (biomol mrna)

Global Entrez Search



Results



291 records

NM_xxxxxx: RefSeqs!

Limit by Title and Database

Entrez Nucleotide

99.3% GenBank (primary data)

`srcdb ddbj/embl/genbank[properties]`

Entrez Trick

0.7% RefSeq (curated data)

`srcdb refseq[properties]`

```
#1: thyroid peroxidase AND human[orgn]                    291
#2: thyroid peroxidase[title] AND human[orgn]             166

#3: #2 AND srcdb refseq[properties]                         5
#4: #2 AND srcdb ddbj/embl/genbank[properties]            161
```

primary data

GenBank Divisions

PRI (27)	Primate
ROD (11)	Rodent
PLN (10)	Plant and Fungal
BCT (8)	Bacterial/Archeal
INV (6)	Invertebrate
VRL (3)	Viral
VRT (4)	Other Vertebrate
MAM (1)	Mammalian
PHG (1)	Phage
SYN (1)	Synthetic
UNA (1)	Unannotated

Traditional

- Direct Submissions (Sequin/Bankit)
- Accurate (~1 error per 10,000 bp)
- Well characterized
- Organized by taxonomy

EST (288)	Expressed Sequence Tag
GSS (98)	Genome Survey Sequence
HTG (61)	High Throughput Genomic
HTC (3)	High Throughput cDNA
STS (3)	Sequence Tagged Site

Bulk

- For sequencing projects
- Batch submissions (ftp/email)
- Inaccurate
- Poorly Characterized
- Organized by sequence type

Limit by Genbank Division

EST Division `gbdiv est[prop]`

Primate Division `gbdiv pri[prop]`

```
#1: thyroid peroxidase AND human[orgn]                    291
#2: thyroid peroxidase[title] AND human[orgn]             166
#3: #2 AND srcdb refseq[properties]                         5
#4: #2 AND srcdb ddbj/embl/genbank[properties]            161

#5: #2 AND gbdiv est[prop]                                 20
#6: #2 AND gbdiv pri[prop]                                141
```

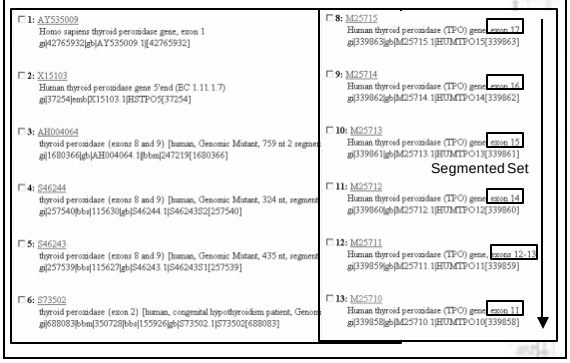
traditional GenBank records

Limit by Biomolecule Type

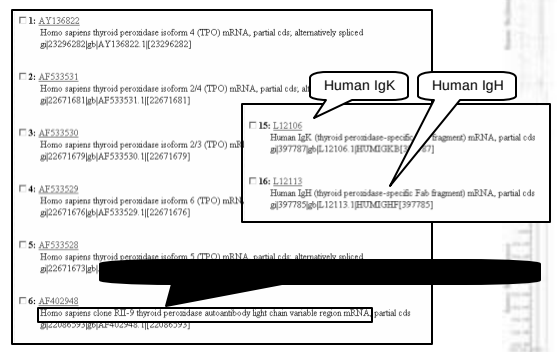
Genomic DNA	`biomol genomic[prop]`
cDNA	`biomol mrna[prop]`

```
#1: thyroid peroxidase AND human[orgn]                  291
#2: thyroid peroxidase[title] AND human[orgn]           166
#3: #2 AND srcdb refseq[properties]                       5
#4: #2 AND srcdb ddbj/embl/genbank[properties]          161
#5: #2 AND gbdiv est[prop]                               20
#6: #2 AND gbdiv pri[prop]                              141
                          genomic DNA
#7: #6 AND biomol genomic[prop]                          25
#8: #6 AND biomol mrna[prop]                            116
                          mRNA / cDNA
```

Genomic Records



cDNA Records



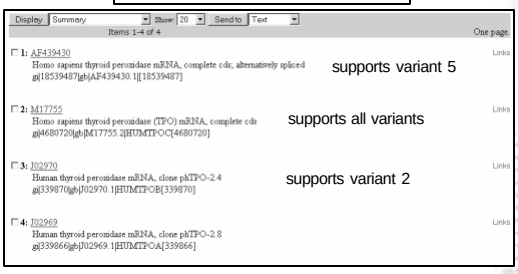
Use NOT

```
thyroid peroxidase[title] AND human[orgn] AND
gbdiv pri[prop] AND biomol mrna[prop] NOT
variable NOT IgK NOT IgH
```

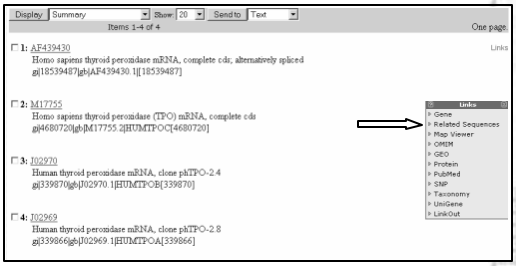
Limit by Protein Name

```
thyroid peroxidase[protein name] AND human[orgn] AND
gbdiv pri[prop] AND biomol mrna[prop]
```

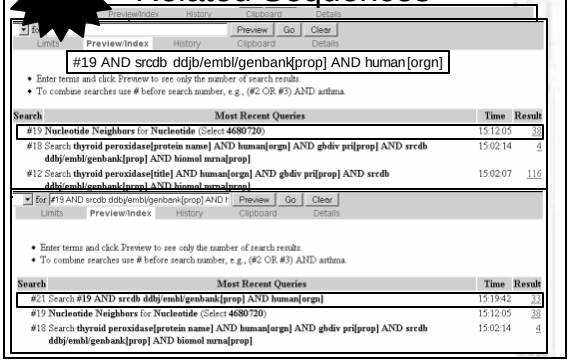
116 records (title) → 4 records (protein name)



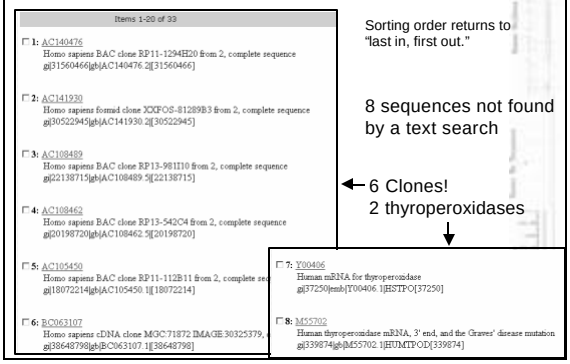
Search by Sequence



Related Sequences



Similar Human GenBank Sequences



Sorting order returns to "last in, first out."

8 sequences not found by a text search

← 6 Clones!
2 thyroperoxidases

How did we do?

Entrez Gene Related Sequences

Related Sequences

Nucleotide		Protein
Genomic	M25715	AAA97517
mRNA	AF439430	AAL74416
mRNA	AF533528	AAN04471
mRNA	AF533529	AAN04472
mRNA	AF533530	AAN04473
mRNA	AF533531	AAN04474
mRNA	AW874082	None
mRNA	AY136822	AAN11302
mRNA	J02969	AAA61215
mRNA	J02970	AAA61216
mRNA	M17755	AAA61217
mRNA	M55702	AAA61218
		AAA61219
mRNA	X17358	CAA35235
mRNA	Y00406	CAA68467
	None	P07202

Entrez Search

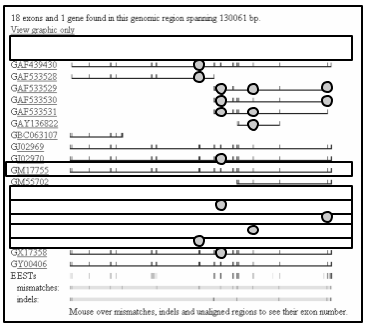
(EST found by gbdiv est[prop])

→ Related Sequences

→ Related Sequences

→ SWISS-PROT, no nucleotide links!

Creating NM Records



Genome annotation

Longest mRNA

NMs must have cDNA support

NM/NP Records in Entrez

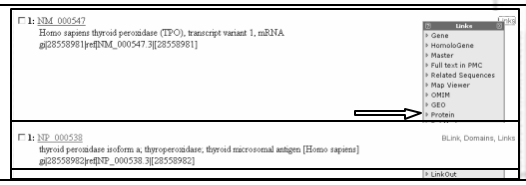
NM_000547: variant 1

```
COMMENT REVIEWED REFSEQ: This record has been curated by NCBI staff.
The reference sequence was derived from M17755.2 and AW874082.1.
On Feb 25, 2003 this sequence version replaced gi:21361188.
```

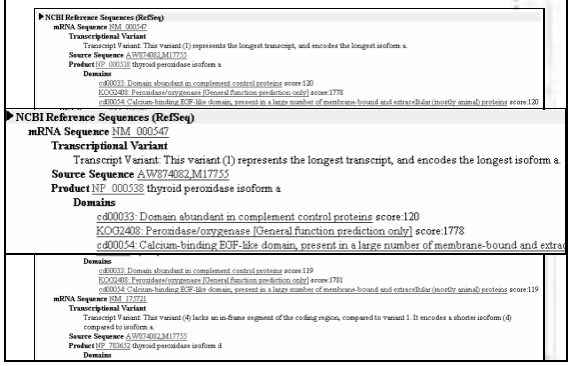
EST that completes 3' end

NM_175719: variant 2

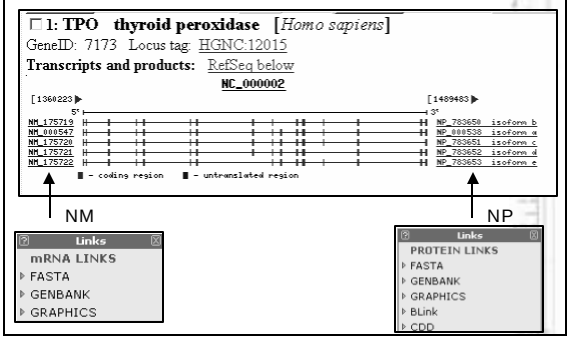
```
COMMENT REVIEWED REFSEQ: This record has been curated by NCBI staff.
The reference sequence was derived from J02970.1, AW874082.1 and M17755.2.
```



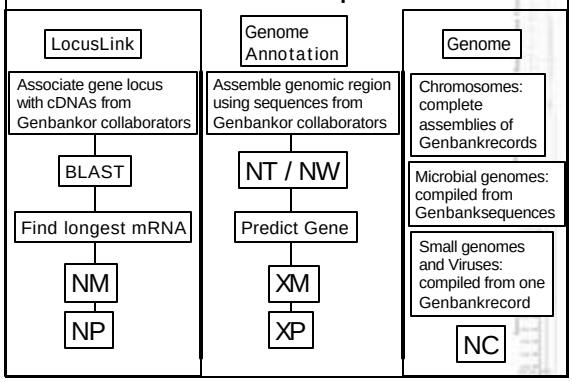
NM/NP Records in Entrez Gene



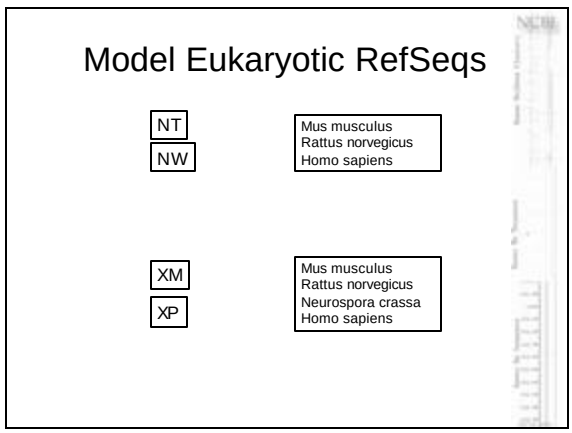
Entrez Gene RefSeq Graphics



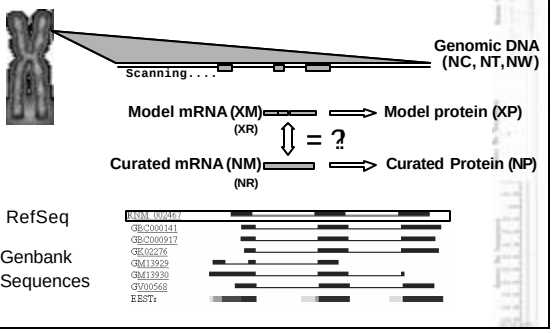
Annotation Pipeline



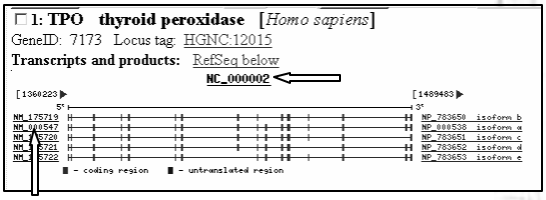
Model Eukaryotic RefSeqs



Creating Transcript Models (XMs)



So how about TPO?

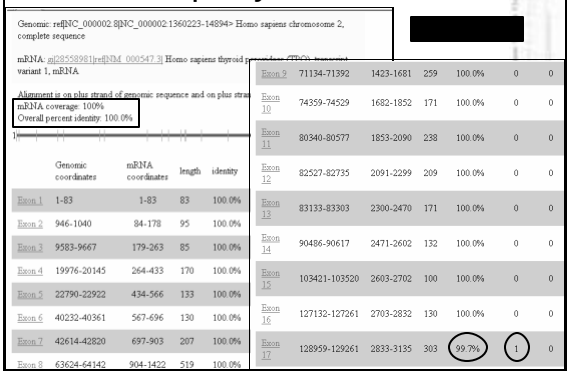


Spidey

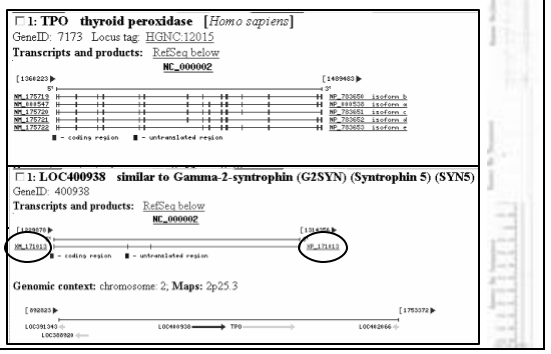
Aligns mRNA to genomic DNA and predicts exon-intron structure

http://www.ncbi.nlm.nih.gov/spidey

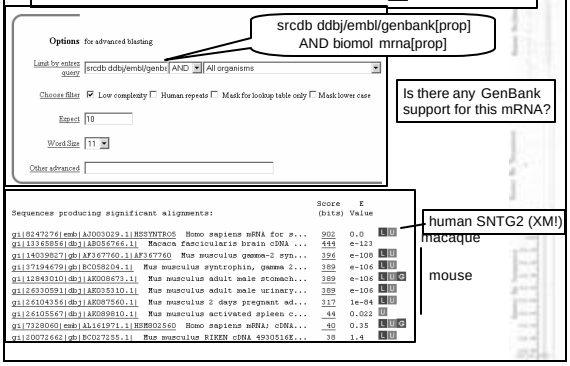
Spidey Results



What about LOC400938?



BLAST Results for XM_171013



Is there any GenBank support for this mRNA?

The Perils of the XM

1: XM_124429

ref|XM_124429.1|[20850420]
This record was removed as a result of standard genome annotation processing. See the genome build documentation at http://www.ncbi.nlm.nih.gov/genome/guide/build.html for further information, or contact info@ncbi.nlm.nih.gov.

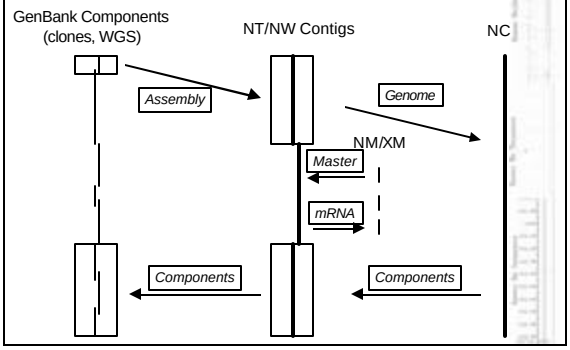
XM records are models based *only* on genomic sequence, and are subject to revision or removal with each new build of that genome.

BLAST the XM against the RefSeq database to look for a replacement:

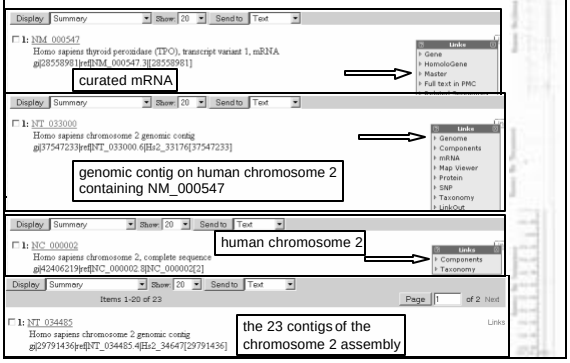
Query= gi|20850420|ref|XM_124429.1|
Mus musculus expressed sequence AA553001 (AA553001), mRNA

gi|19527087|ref|NM_133873.1|
Mus musculus DNA segment, Chr 4, Wayne State University 114, expressed (D4Wsu114e), mRNA Length=1898
Score = 3701.55 bits (1867), Expect = 0
Identities = 1870/1871 (99%), Gaps = 0/1871 (0%) Strand=Plus/Plus

Genome Annotation in Entrez



Genome Annotation Links

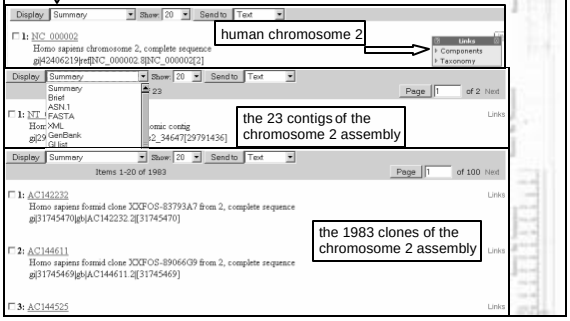


Annotation Links

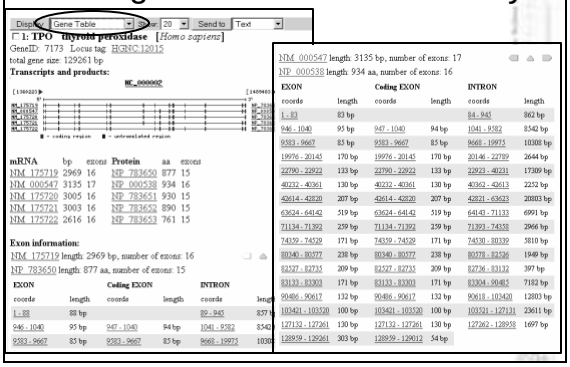
Entrez Trick

Problem

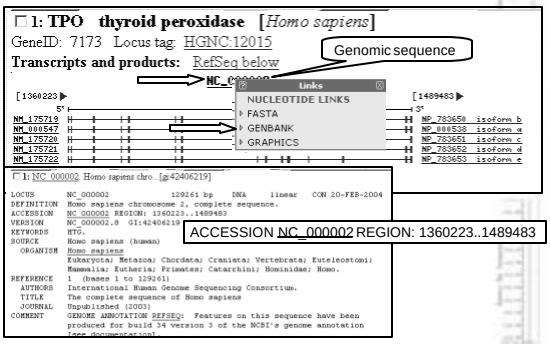
Download all clones used in the assembly of human chromosome 2



Getting an Annotation Summary



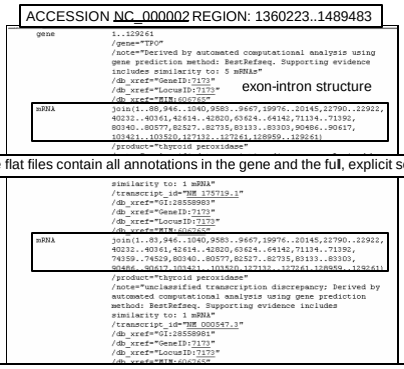
Getting the Annotation Details



ACCESSION NC_000002 REGION: 1360223..1489483

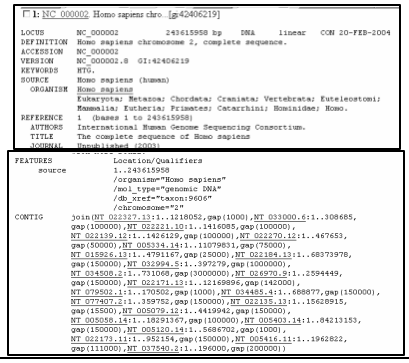
Getting the Annotation Details

ACCESSION NC_000002 REGION: 1360223..1489483



These flat files contain all annotations in the gene and the full, explicit sequence

Getting the Assembly Details

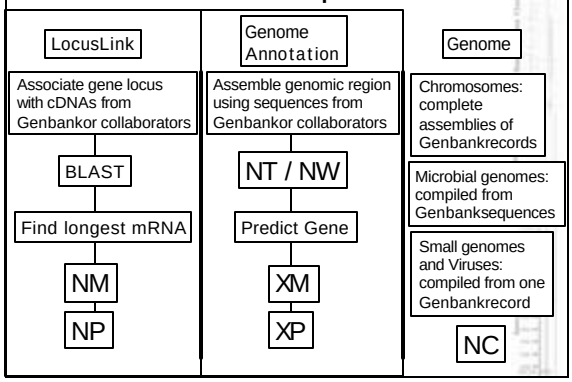


1: NC_000002 Homo sapiens chro... [gi:42406219]

```
LOCUS       NC_000002          243615958 bp    DNA     linear   CON 20-FEB-2004
DEFINITION  Homo sapiens chromosome 2, complete sequence.
ACCESSION   NC_000002
VERSION     NC_000002.8  GI:42406219
KEYWORDS    HTG.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 243615958)
  AUTHORS   International Human Genome Sequencing Consortium.
  TITLE     The complete sequence of Homo sapiens
  JOURNAL   Unpublished (2001)
```

```
FEATURES             Location/Qualifiers
     source          1..243615958
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /chromosome="2"
CONTIG       join(NT_022327.13:1..1218052,gap(1000),NT_033000.6:1..308685,
             gap(100000),NT_022221.10:1..1416085,gap(100000),
             NT_022139.12:1..1426129,gap(100000),NT_022270.12:1..467653,
             gap(50000),NT_005334.14:1..1079831,gap(75000),
             NT_015926.13:1..4791167,gap(25000),NT_022184.13:1..68373978,
             gap(150000),NT_032994.5:1..397279,gap(1000000),
             NT_034508.2:1..731069,gap(3000000),NT_026970.9:1..2594449,
             gap(150000),NT_022171.13:1..12169896,gap(142000),
             NT_079502.1:1..170502,gap(1000),NT_034485.4:1..688877,gap(150000),
             NT_077407.2:1..359782,gap(150000),NT_022135.13:1..15628915,
             gap(15500),NT_005079.12:1..4419942,gap(150000),
             NT_005058.14:1..16291367,gap(100000),NT_005403.14:1..8421153,
             gap(150000),NT_005120.14:1..5686702,gap(1000),
             NT_022173.11:1..952154,gap(150000),NT_005416.11:1..1962822,
             gap(111000),NT_037540.2:1..196000,gap(200000))
```

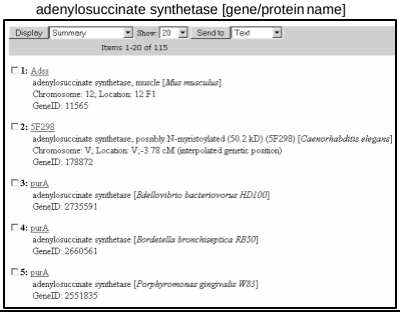
Genome Pipeline



Entrez Gene and Genome

Entrez Gene now includes all organisms in Entrez Genome

adenylosuccinate synthetase [gene/protein name]



Display Summary Show: 20 Send to Text

Items 1-20 of 115

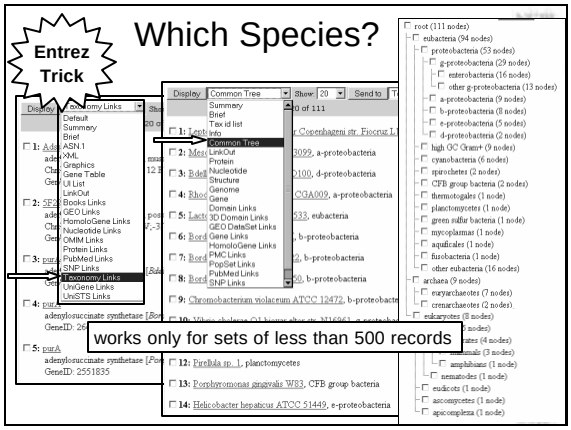
1: Adss
adenylosuccinate synthetase, muscle [*Mus musculus*]
Chromosome: 12; Location: 12 F1
GeneID: 11565

2: 5F298
adenylosuccinate synthetase, possibly N-myristoylated (50.2 kD) (5F298) [*Caenorhabditis elegans*]
Chromosome: V; Location: V; -3.78 cM (interpolated genetic position)
GeneID: 178872

3: purA
adenylosuccinate synthetase [*Bdellovibrio bacteriovorus HD100*]
GeneID: 2735591

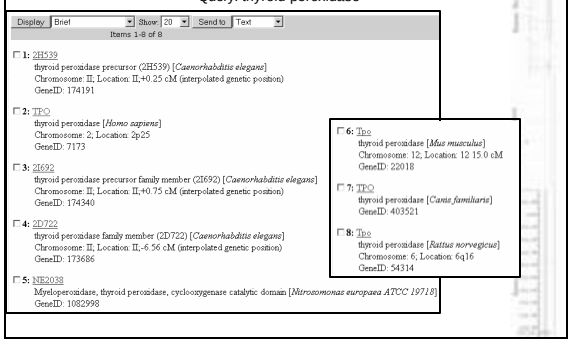
4: purA
adenylosuccinate synthetase [*Bordetella bronchiseptica RB50*]
GeneID: 2660561

5: purA
adenylosuccinate synthetase [*Porphyromonas gingivalis W83*]
GeneID: 2551835



Finding Genes by Text

Query: thyroid peroxidase



Searching Entrez Gene

Gene symbol: human thyroid peroxidase (TPO)

tpo [sym] AND human [organism]

Protein name: topoisomerase genes from Archaea

topoisomerase[gene/protein name] AND archaea [organism]

Chromosome and Links: genes on human chromosome 2 with OMIM links

2 [chromosome] AND gene omim [filter] AND human [organism]

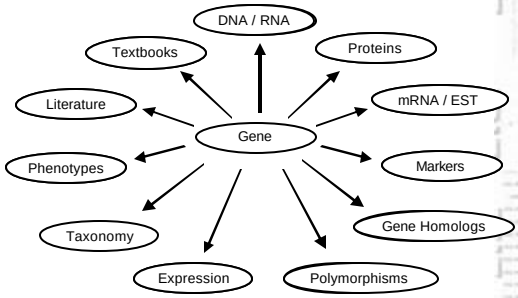
RefSeq status and variants: Reviewed RefSeqs with transcript variants

srcdb refseq reviewed[prop] AND has transcript variants[prop]

Disease and Gene Ontology: Membrane proteins linked to cancer

integral to plasma membrane[gene ontology] AND cancer [dis]

Gene Links in Entrez



Scripting Gene Searches

E-utilities

Conduct Entrez searching by posting URLs

Find human genes with disease alleles that have been mapped uniquely

ESearch(gene)	**human [organism] AND gene omim[filter]**	#1: Gene UIDs
ELink(gene, OMIM)	#1: Gene UIDs →	#2: OMIM UIDs
ESearch(OMIM)	#2 **AND allelic variants[prop] AND has asterisk[prop]** →	#3: OMIM UIDs
ELink(OMIM Gene)	#3: OMIM UIDs →	#4: Gene UIDs
ESummary	#4: Gene UIDs →	Summaries

Gene, LocusLink, and the Future

- Gene will eventually replace LocusLink
- Many LocusLink functions are now in the Entrez Links menu
- As much as possible, LocusLink IDs are equal to the corresponding Gene UID
- Currently there is no FTP site for Entrez Gene
- Therefore, we will maintain the LL_tmpl file and associated files until Entrez Gene has FTP files

Beyond RefSeq

If your organism does not have RefSeqs…

- UniGene : gene-based clusters of cDNAs and ESTs
- UniSTS : STS markers
- WGS sequences in Entrez Nucleotide (wgs[prop])
- Trace Archive

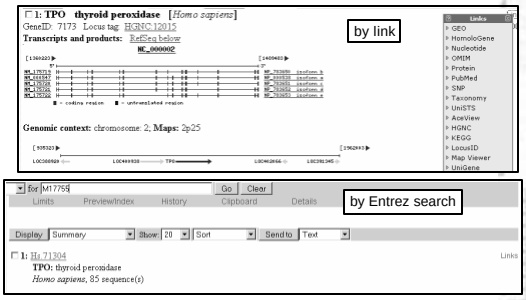
UniGene: Clustering Expressed Sequences

- Records are clusters of mRNAs and ESTs that ideally represent single genes
- Records are created automatically by a modified BLAST algorithm
- UniGene provides a means to identify an EST or unannotated mRNA

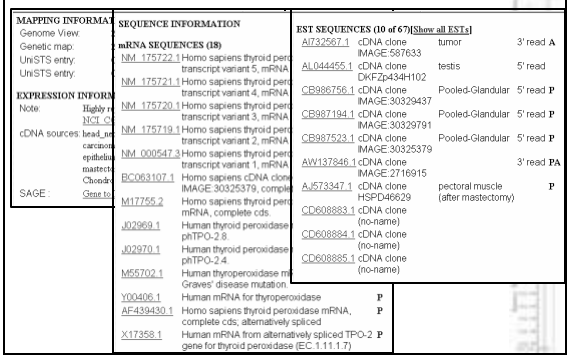
Organisms in UniGene

Species	UniGene Entries
Chordata	
Mammalia	
Bos taurus	23,572
Canis familiaris	3,905
Homo sapiens	105,680
Mus musculus	79,863
Rattus norvegicus	39,874
Sus scrofa	20,426
Aves	
Gallus gallus	12,220
Amphibia	
Xenopus laevis	22,129
Xenopus tropicalis	17,102
Actinopterygii	
Danio rerio	17,951
Oncorhynchus mykiss	13,458
Oryzias latipes	6,881
Salmo salar	1,033
Ascidiacea	
Ciona intestinalis	13,600
Echinodermata	
Echinoidea	
Strongylocentrotus purpuratus	2,592
Arthropoda	
Insecta	
Anopheles gambiae	14,085
Apis mellifera	4,756
Bombyx mori	2,008
Drosophila melanogaster	12,322
Nematoda	
Chromadorea	
Caenorhabditis elegans	15,655
Platyhelminthes	
Trematoda	
Schistosoma mansoni	1,185
Embryophyta	
Bryopsida	
Physcomitrella patens	6,946
Coniferopsida	
Pinus taeda	8,669
Eudicotyledons	
Arabidopsis thaliana	19,666
Glycine max	12,567
Helianthus annuus	1,897
Lactuca sativa	9,629
Lycopersicon esculentum	3,658
Medicago truncatula	5,396
Populus tremula x Populus tremuloides	2,711
Solanum tuberosum	5,476
Vitis vinifera	11,832
Liliopsida	
Hordeum vulgare	11,980
Oryza sativa	25,089
Saccharum officinarum	4,771
Sorghum bicolor	7,075
Triticum aestivum	24,543
Zea mays	12,941
Chlorophyta	
Chlorophyceae	
Chlamydomonas reinhardtii	5,874
Mycetozoa	
Dictyostelida	
Dictyostelium discoideum	3,853
Apicomplexa	
Coccidia	
Toxoplasma gondii	6,001

Finding UniGene Clusters



UniGene Cluster for TPO



Searching Entrez UniGene

Nucleotide: Non-genomic sequences from the PLN division of Genbank

gbdiv_pln [properties] NOT biomol_genomic [properties]

mRNA: rat clusters that have at least one mRNA

rat [organism] NOT 0 [mrna count]

Chromosome: e

Snp [SNP Class] AND 1 [Map Weight] AND mouse [organism]

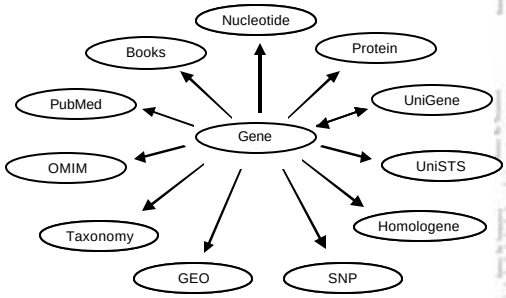
UniSTS: Markers on the Genethon map of human chromosome 12

Genethon [Map Name] AND human [organism] AND 12 [chromosome]

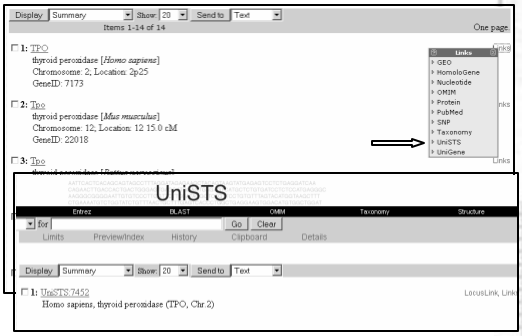
Structure: Structures of bacterial kinases with resolutions below 2 Å

Bacteria [organism] AND kinase AND 000.00:002.00 [resolution]

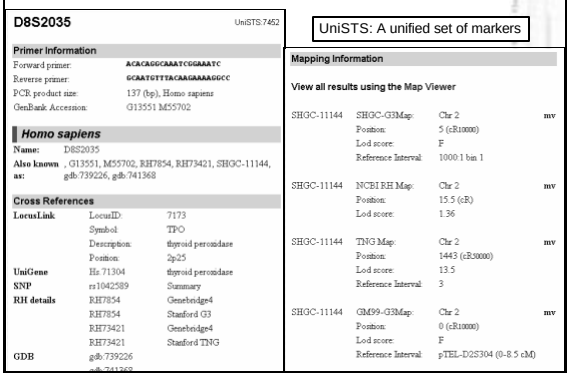
UniGene Links in Entrez



UniSTS

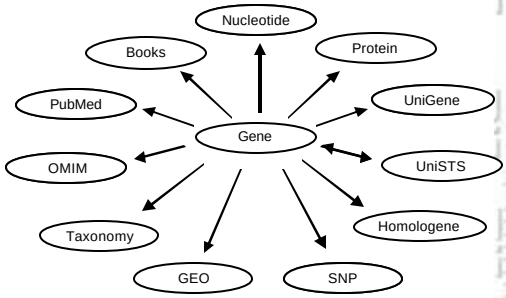


UniSTS Record



UniSTS: A unified set of markers

UniSTS Links in Entrez

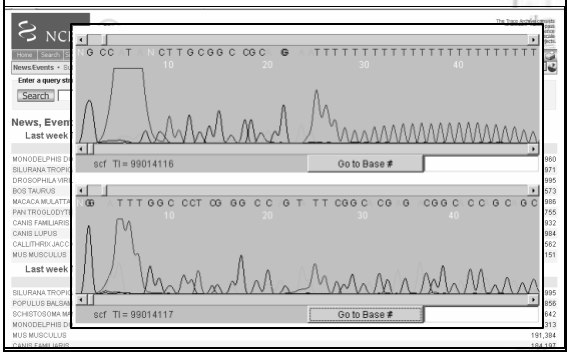


Whole Genome Shotgun

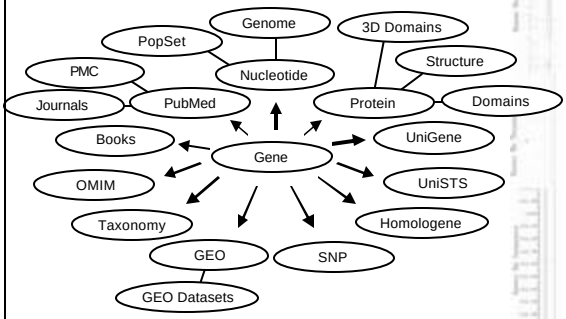
wgs master[properties] wgs contig[properties]

```
LOCUS       CAAA01000000          224713 rc     DNA     linear   ROD 15-MAY-2002
DEFINITION  Mus musculus whole genome shotgun sequencing project.
ACCESSION   CAAA00000000
VERSION     CAAA00000000.1  GI:20800445
KEYWORDS    WGS.
SOURCE      house mouse.
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
REFERENCE   1
  AUTHORS   The Mouse Genome Sequencing,Consortium.
  TITLE     Direct Submission
  JOURNAL   Submitted (01-MAY-2002) The Mouse Genome Sequencing Consortium,
            http://mouse.ensembl.org/
COMMENT     The Mus musculus whole genome shotgun (WGS) project has the project
            accession CAAA00000000.  This version of the project (01) has the
            accession number CAAA01000000, and consists of sequences
            CAAA01000001-CAAA01224713.
FEATURES             Location/Qualifiers
     source          1..224713
                     /organism="Mus musculus"
                     /db_xref="taxon:10090"
WGS         CAAA01000001-CAAA01224713
//
```

Trace Archive



Linking Beyond Genes



Sequence Polymorphisms

SNP

General Polymorphisms

- Primary database of submitted SNPs
- Curated database of reference SNPs
- Contains more than just SNPs:
  - True SNPs
  - MNP (multiple nucleotide)
  - Insertions
  - Deletions
  - Microsatellites
  - Mixed
  - No variation (constant)

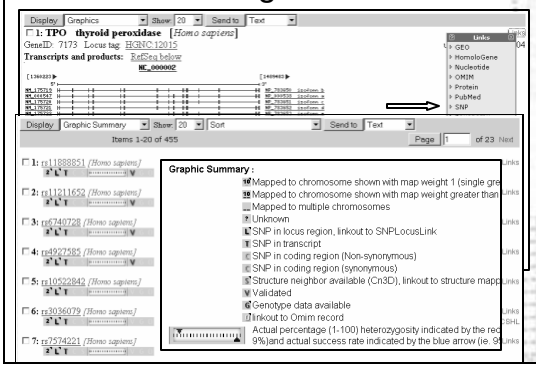
OMIM

Human Phenotypes

- Clinical literature database
- Curated at Johns Hopkins Univ
- Links human genes and genetic disorders to human disease
- Lists allelic variants that have clinical consequences

Variations in SNP are not necessarily in OMIM, and vice versa!

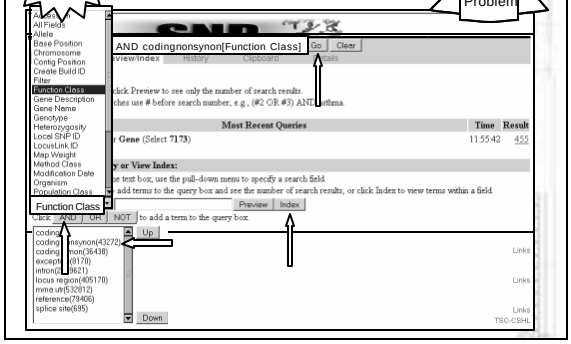
Linking to SNP



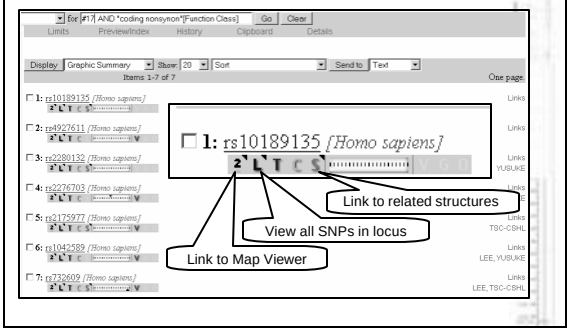
Find Non-synonymous SNPs

Entrez Trick

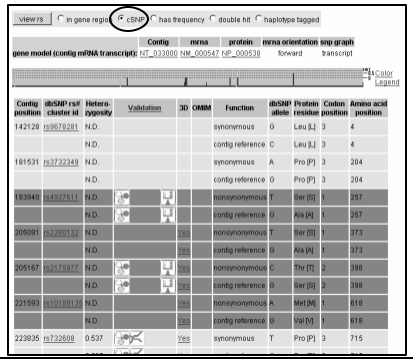
Problem



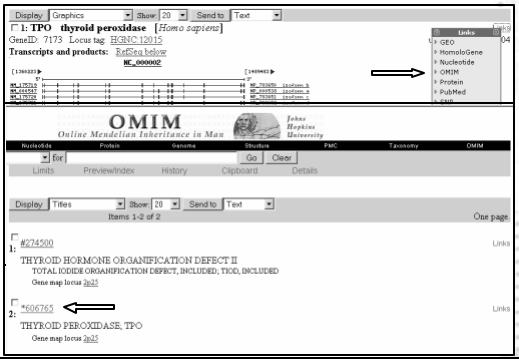
Non-synonymous TPO SNPs



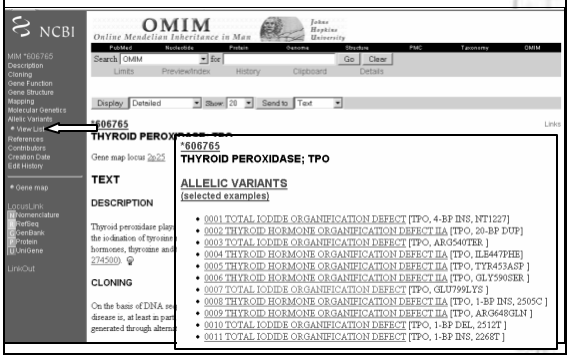
Locus View of SNPs



Links to OMIM



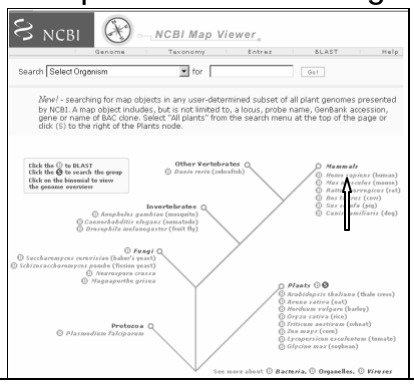
OMIM Record



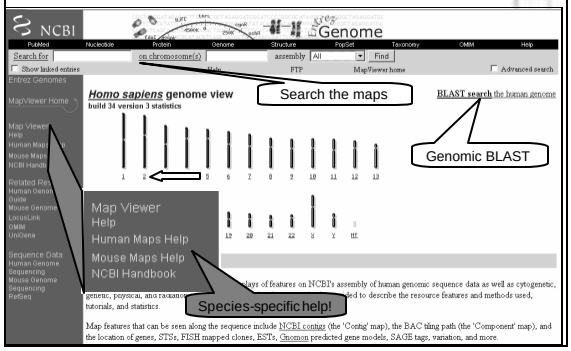
NCBI Map Viewer

- Map Viewer Home Page
  - Shows all supported organisms
  - Provides links to genomic BLAST
- Genome Overview Page
  - Provides links to individual chromosomes
  - Shows hits on a genome graphically
- Chromosome Viewing Page
  - Allows interactive views of annotation details
  - Provides numerous maps unique to each genome

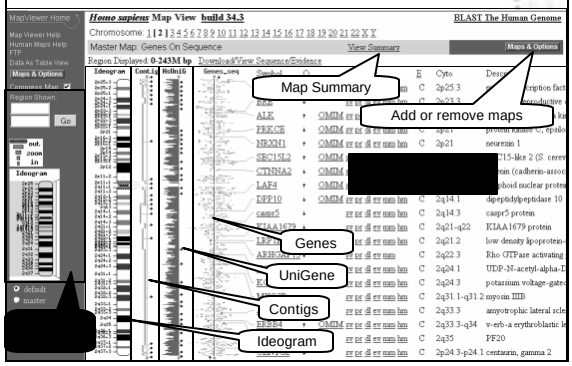
Map Viewer Home Page



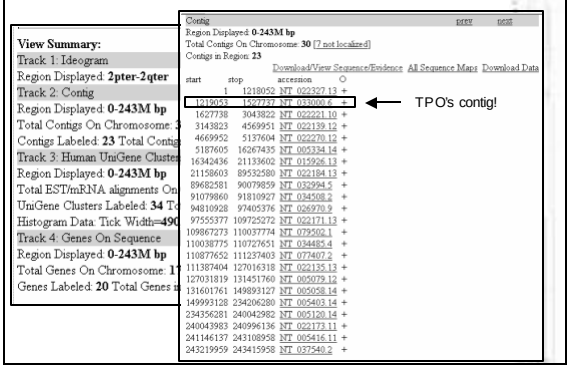
Genome Overview Page



Chromosome Viewing Page



Map Summary



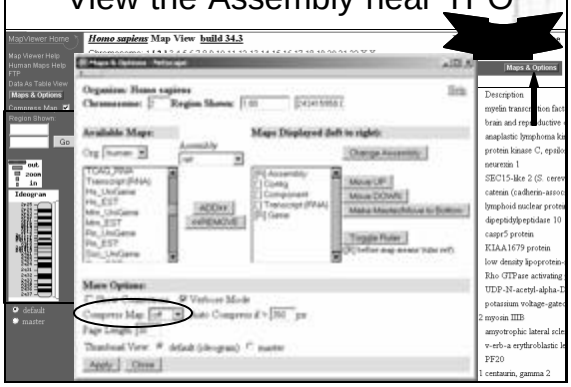
Map Content

Map content varies greatly by species!

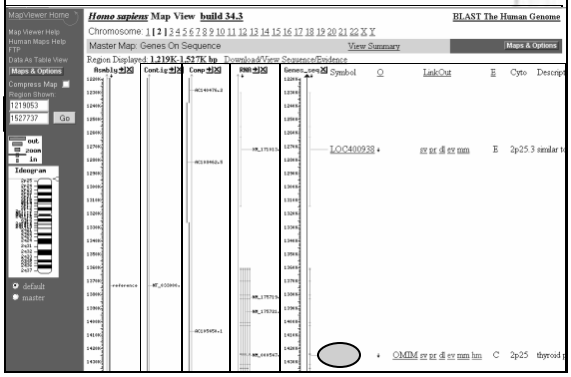
- Sequence Maps
  - Core assembly
  - Annotation evidence
  - Clones & Markers
  - Polymorphisms
  - Links & Features
- Genetic Maps
  - Cytogenetic maps
  - Linkage maps
  - Radiation hybrid maps

Assembly
Contig
Component
Transcript
Gene

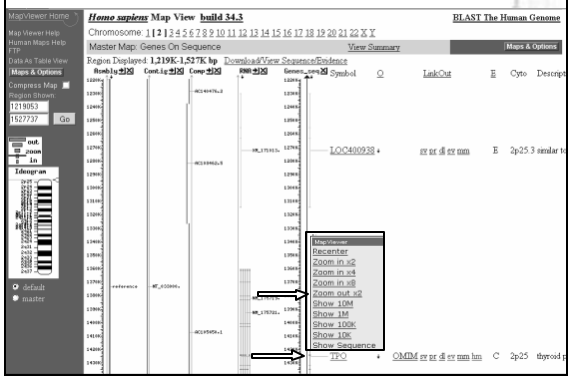
View the Assembly near TPO



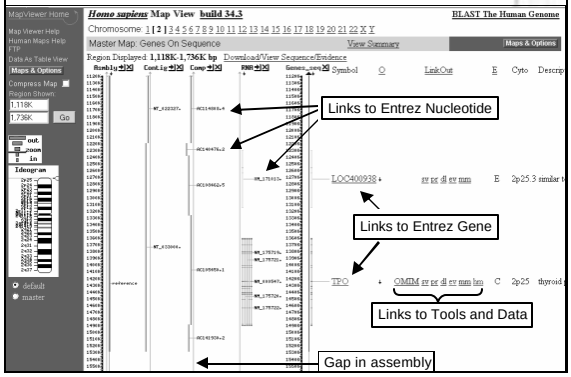
Assembly of Chr. 2



Zooming



View of TPO



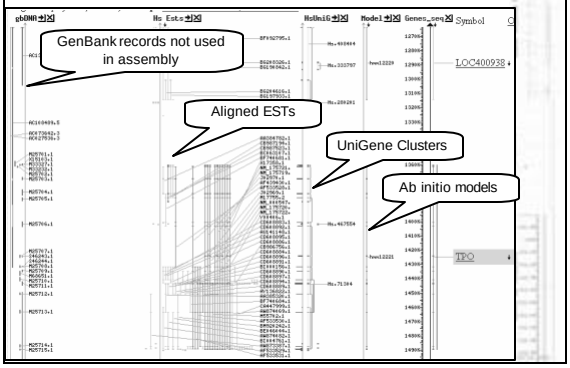
Map Content

Map content varies greatly by species!

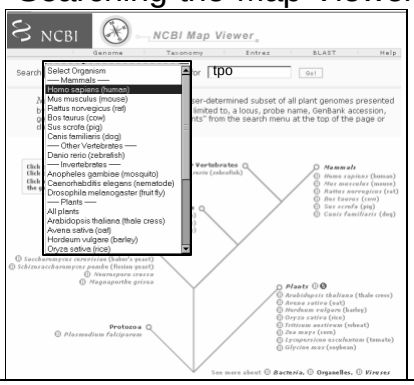
- Sequence Maps
  - Core assembly
  - Annotation evidence
  - Clones & Markers
  - Polymorphisms
  - Links & Features
- Genetic Maps
  - Cytogenetic maps
  - Linkage maps
  - Radiation hybrid maps

Ab initio (model)
GenBank DNA
EST
UniGene
Gene

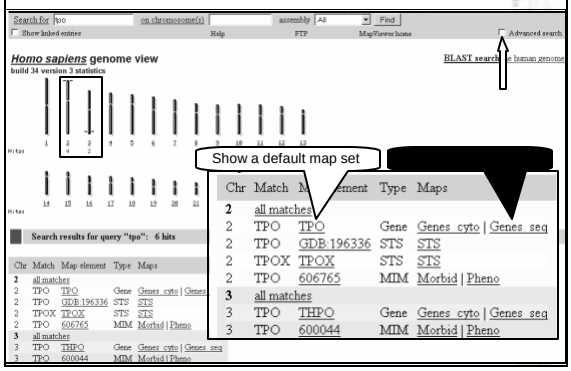
Annotation Evidence



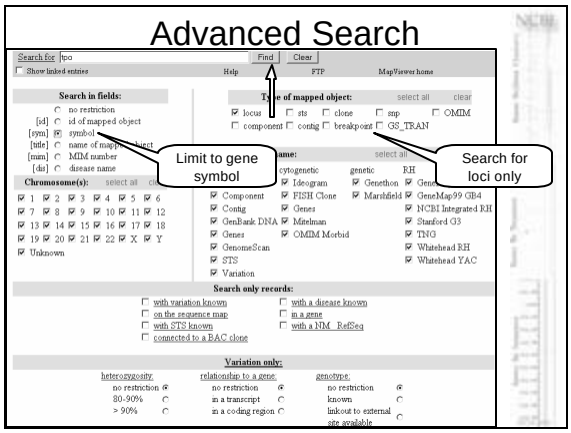
Searching the Map Viewer



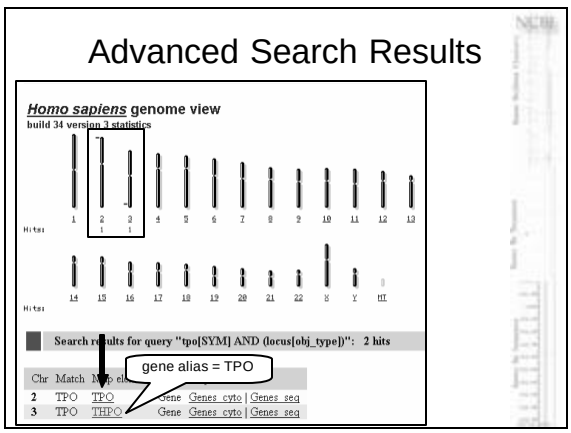
Results



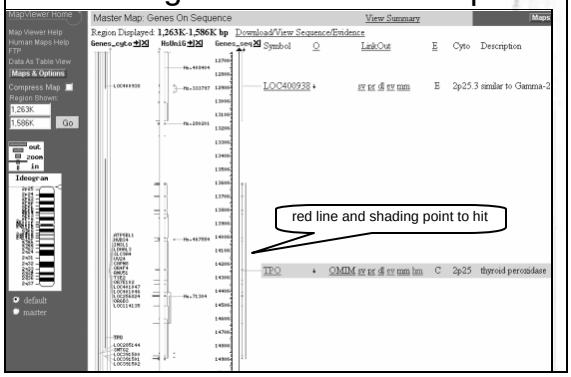
Advanced Search



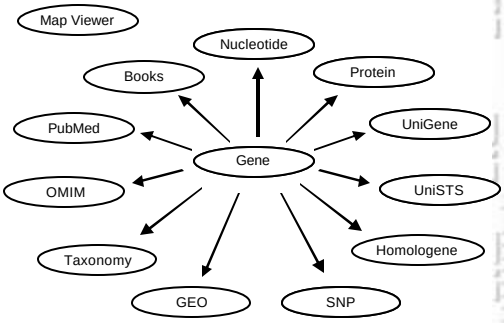
Advanced Search Results



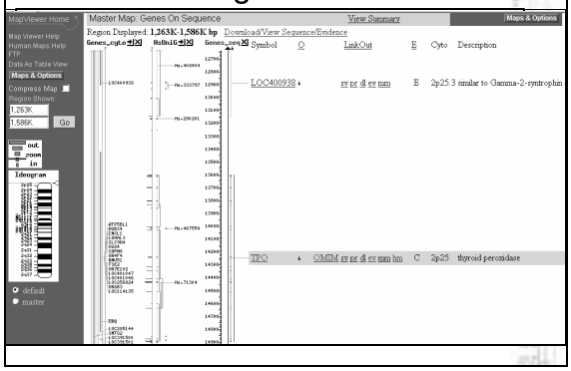
Viewing Results on the Maps



Linking to the Map Viewer

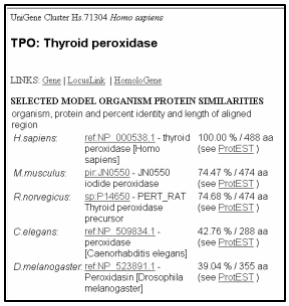


Linking from Gene



Gene Homologs in Entrez

ProtEST in UniGene: Linking Protein Homologs


UniGene Cluster Hs.71304 *Homo sapiens*

TPO: Thyroid peroxidase

LINKS: Gene | LocusLink | HomoloGene

SELECTED MODEL ORGANISM PROTEIN SIMILARITIES
organism, protein and percent identity and length of aligned region

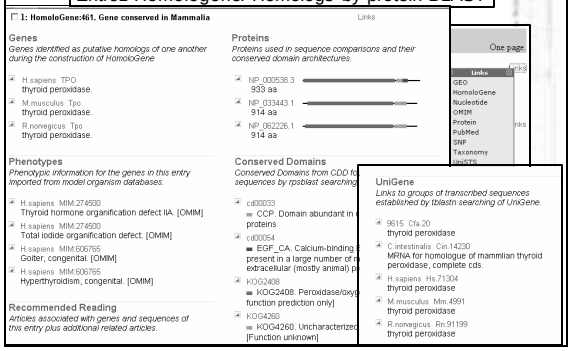
H.sapiens:	ref.NP_000538.1 - thyroid peroxidase [Homo sapiens]	100.00 % / 488 aa (see ProtEST)
M.musculus:	pir.JN0550 - JN0550 iodide peroxidase	74.47 % / 474 aa (see ProtEST)
R.norvegicus:	sp.P14650 - PERT_RAT Thyroid peroxidase precursor	74.68 % / 474 aa (see ProtEST)
C.elegans:	ref.NP_509834.1 - peroxidase [Caenorhabditis elegans]	42.76 % / 288 aa (see ProtEST)
D.melanogaster:	ref.NP_523891.1 - Peroxidasin [Drosophila melanogaster]	39.04 % / 355 aa (see ProtEST)

ProtEST
blastx of UniGene cluster sequences against curated proteins from 8 organisms:

- H. sapiens
- M. musculus
- R. norvegicus
- D. melanogaster
- C. elegans
- S. cerevisiae
- A. thaliana
- E. coli

Gene Homologs in Entrez

Entrez Homologene: Homologs by protein BLAST


1: HomoloGene:461. Gene conserved in Mammalia

Genes
Genes identified as putative homologs of one another during the construction of HomoloGene.

- H.sapiens TPO thyroid peroxidase.
- M.musculus Tpo thyroid peroxidase.
- R.norvegicus Tpo thyroid peroxidase.

Proteins
Proteins used in sequence comparisons and their conserved domain architectures.

- NP_000538.3 933 aa
- NP_033443.1 914 aa
- NP_062226.1 914 aa

Phenotypes
Phenotypic information for the genes in this entry imported from model organism databases.

- H.sapiens MIM:274500 Thyroid hormone organification defect IIA. [OMIM]
- H.sapiens MIM:274500 Total iodide organification defect. [OMIM]
- H.sapiens MIM:606765 Goiter, congenital. [OMIM]
- H.sapiens MIM:606765 Hyperthyroidism, congenital. [OMIM]

Recommended Reading
Articles associated with genes and sequences of this entry plus additional related articles.

Conserved Domains
Conserved Domains from CDD fo sequences by rpsblast searching

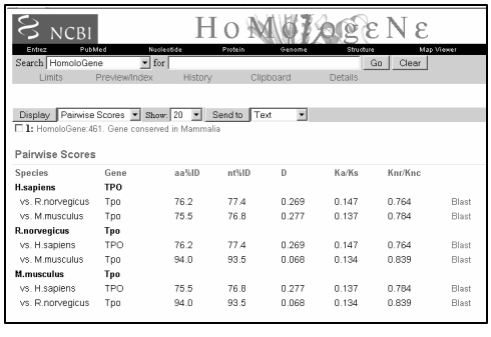
- cd00033 CCP. Domain abundant in proteins
- cd00054 EGF_CA. Calcium-binding present in a large number of extracellular (mostly animal) p
- KOG2408. Peroxidase/oxy function prediction only]
- KOG4260 KOG4260. Uncharacterized [Function unknown]

Links
One page
GEO
HomoloGene
Nucleotide
OMIM
Protein
PubMed
SNP
Taxonomy

UniGene
Links to groups of transcribed sequences established by tblastn searching of UniGene.

- 9615 Cfa.20 thyroid peroxidase
- C.intestinalis Cin.14230 MRNA for homologue of mammlian thyroid peroxidase, complete cds.
- H.sapiens Hs.71304 thyroid peroxidase
- M.musculus Mm.4991 thyroid peroxidase
- R.norvegicus Rn.91199 thyroid peroxidase

Pairwise Scores


NCBI HomoloGene

Entrez | PubMed | Nucleotide | Protein | Genome | Structure | Map Viewer

Search HomoloGene for | Go | Clear

Limits | Preview/Index | History | Clipboard | Details

Display Pairwise Scores | Show: 20 | Send to: Text

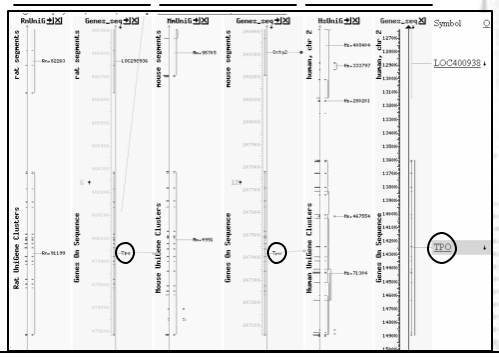
1: HomoloGene:461. Gene conserved in Mammalia

Pairwise Scores

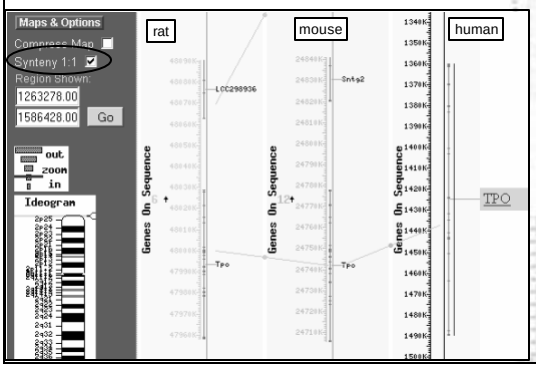
Species	Gene	aa%ID	nt%ID	D	Ka/Ks	Kmr/Knc	
H.sapiens	**TPO**						
vs. R.norvegicus	Tpo	76.2	77.4	0.269	0.147	0.764	Blast
vs. M.musculus	Tpo	75.5	76.8	0.277	0.137	0.784	Blast
R.norvegicus	**Tpo**						
vs. H.sapiens	TPO	76.2	77.4	0.269	0.147	0.764	Blast
vs. M.musculus	Tpo	94.0	93.5	0.068	0.134	0.839	Blast
M.musculus	**Tpo**						
vs. H.sapiens	TPO	75.5	76.8	0.277	0.137	0.784	Blast
vs. R.norvegicus	Tpo	94.0	93.5	0.068	0.134	0.839	Blast

Gene Homologs in Map Viewer

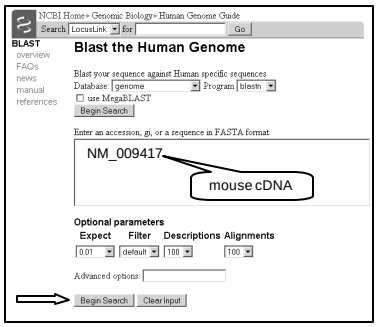
Rat – Mouse - Human



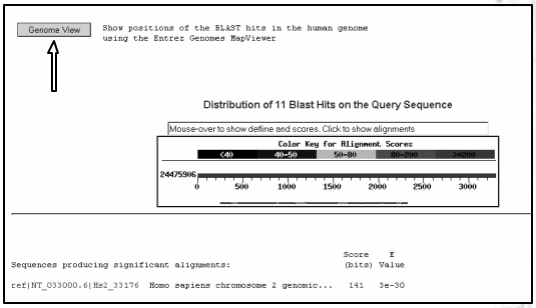
A Closer Look…



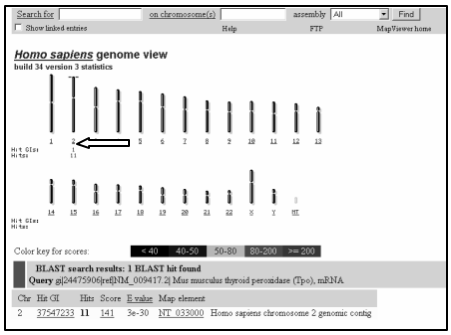
Genomic BLAST



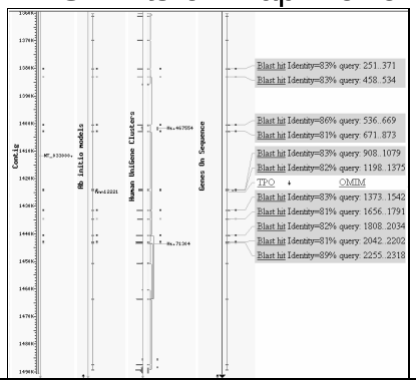
Results Summary



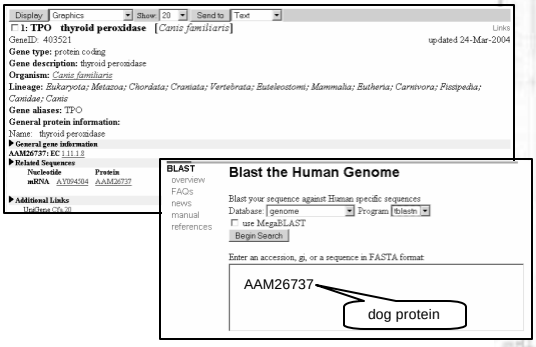
Results on Genome Overview



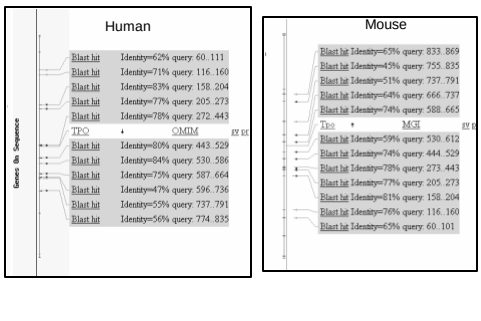
BLAST Hits on Map Viewer



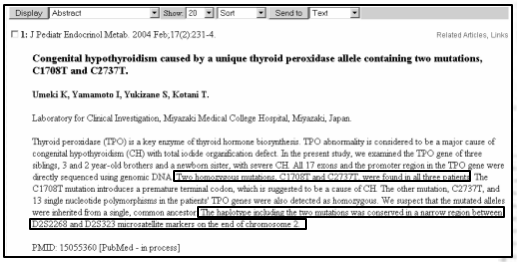
Comparing the Dog cDNA



tBLASTn Results



Find a Locus Between Markers



Display | Abstract | Show: 20 | Sort | Send to | Text

1: J Pediatr Endocrinol Metab. 2004 Feb;17(2):231-4. Related Articles, Links

Congenital hypothyroidism caused by a unique thyroid peroxidase allele containing two mutations, C1708T and C2737T.

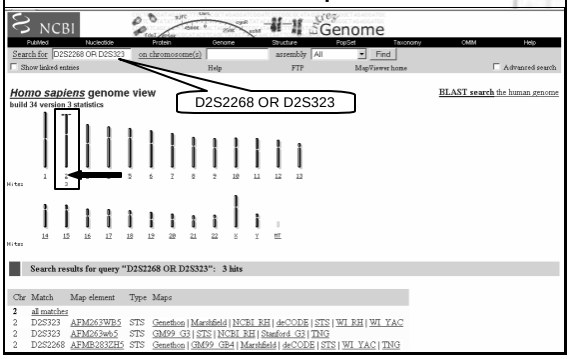
Umeki K, Yamamoto I, Yukizane S, Kotani T.

Laboratory for Clinical Investigation, Miyazaki Medical College Hospital, Miyazaki, Japan.

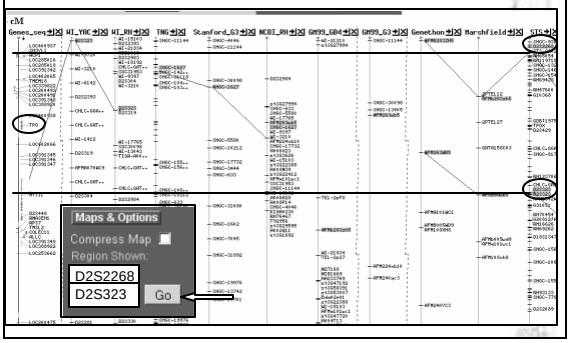
Thyroid peroxidase (TPO) is a key enzyme of thyroid hormone biosynthesis. TPO abnormality is considered to be a major cause of congenital hypothyroidism (CH) with total iodide organification defect. In the present study, we examined the TPO gene of three siblings, 3 and 2 year-old brothers and a newborn sister, with severe CH. All 17 exons and the promoter region in the TPO gene were directly sequenced using genomic DNA. Two homozygous mutations, C1708T and C2737T, were found in all three patients. The C1708T mutation introduces a premature terminal codon, which is suggested to be a cause of CH. The other mutation, C2737T, and 13 single nucleotide polymorphisms in the patients' TPO genes were also detected as homozygous. We suspect that the mutated alleles were inherited from a single, common ancestor. The haplotype including the two mutations was conserved in a narrow region between D2S2268 and D2S323 microsatellite markers on the end of chromosome 2.

PMID: 15055360 [PubMed - in process]

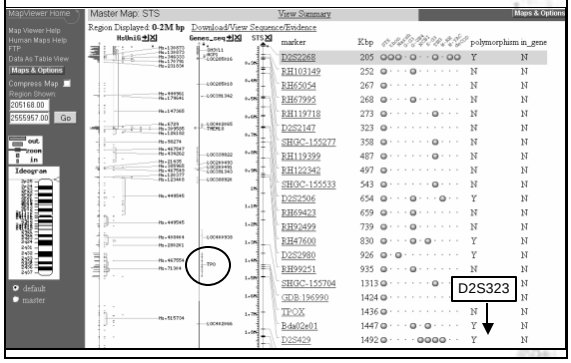
Markers in Map Viewer



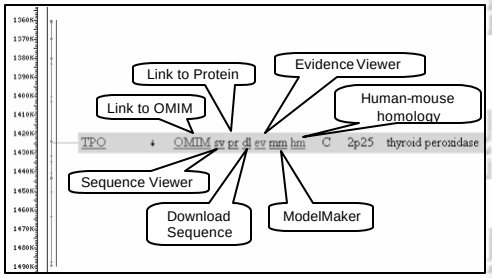
Targeting a Region



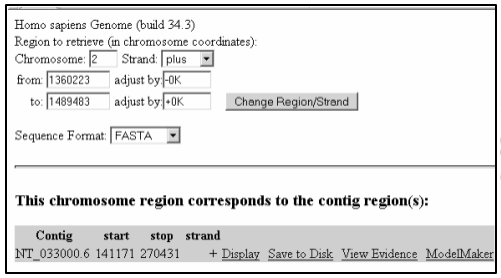
D2S2268 – D2S323



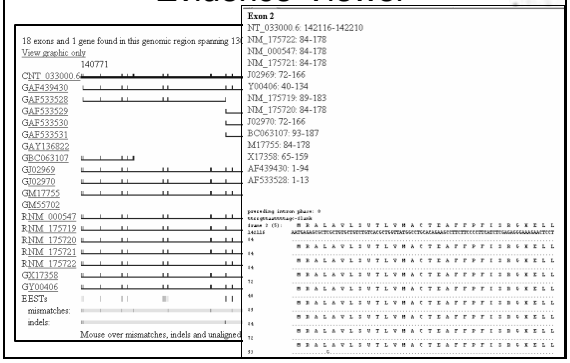
Map Viewer Analysis Tools



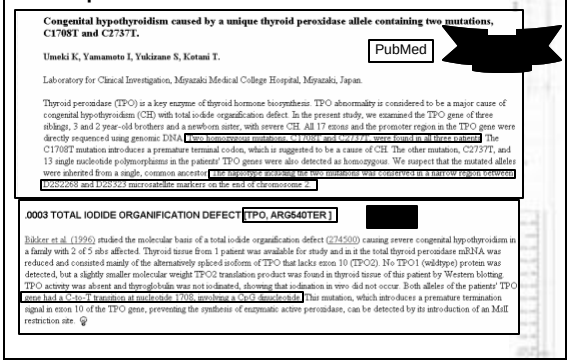
Download Link: dl



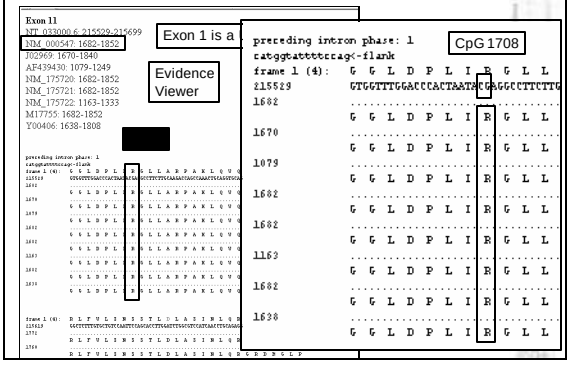
Evidence Viewer



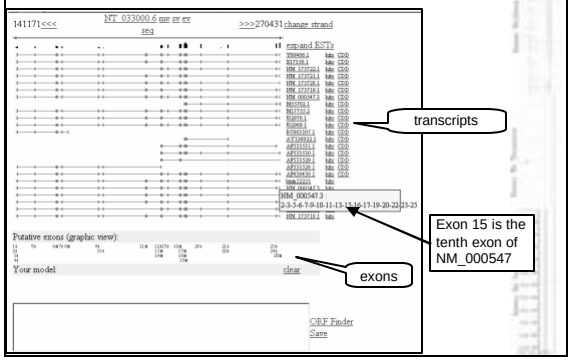
Explore a Disease Allele: C1708T



Locate C1708 / R540



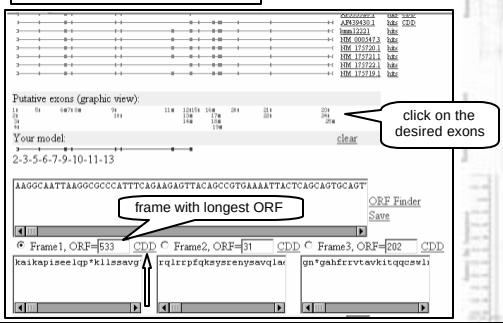
ModelMaker



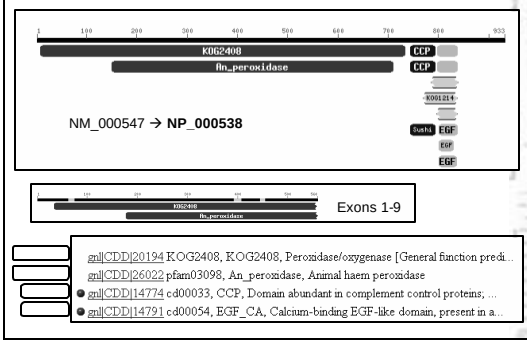
Create a Model with Exons 1-9

NM_000547 = 2-3-5-6-7-9-10-11-13-15-16-17-19-20-22-23-25

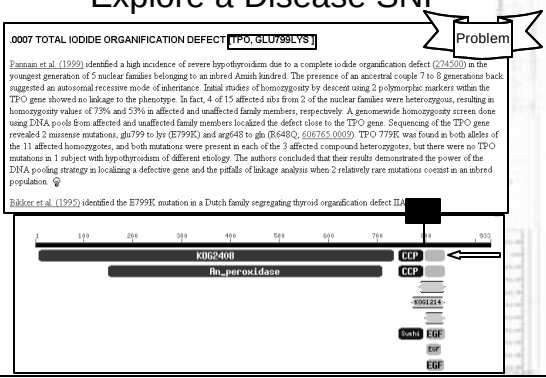
Desired model = 2-3-5-6-7-9-10-11-13



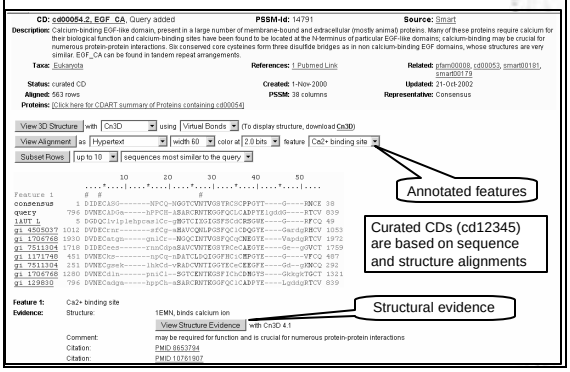
Compare Conserved Domains



Explore a Disease SNP

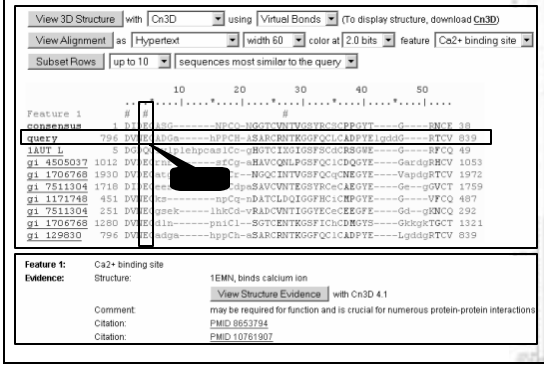


Curated CD Record



Curated CD Record

Cn3D



For More Information…

• **Eric Sayers**	**sayers@ncbi.nlm.nih.gov**
• **General Help**	**info@ncbi.nlm.nih.gov**
• **BLAST**	**blast-help@ncbi.nlm.nih.gov**

The (free!) NCBI Newsletter

http://www.ncbi.nih.gov/About/newsletter.html

The NCBI Handbook

Follow the link from the NCBI Home Page

The NCBI Education Page

http://www.ncbi.nih.gov/Education/index.html

Come See Us!

Booth #12

Poster #233